Multi-View Domain Adaptation for Nighttime Aerial Tracking

Haoyang Li[1], Guangze Zheng[2], Sihang Li[1], Junjie Ye[1], Changhong Fu[1,*]

***Abstract*— Daytime-Nighttime domain adaptation (DNDA) has significantly extended intelligent visual applications of unmanned aerial vehicles (UAVs). However, existing method that merely relies on single-view information neglects the significant differences in viewpoint and motion pattern disparities across different views, resulting in limited performance and robustness in adapting to aerial view variations. Moreover, shadow occlusion, uneven lighting distribution, and disruptive noise exacerbate multi-view feature differences in the nighttime, which may leads to missed targets or tracking failures. To address these issues, this work presents a domain adaptation framework with aerial multi-view source domains for nighttime aerial tracking named MVDANT. Specifically, a nighttime tracking training strategy fusing with multi-view knowledge is proposed. Multi-view domain adaptation is employed to narrow the huge gap between daytime and nighttime scenarios by capturing images from multiple views in daytime scenarios. Additionally, an innovative self-attention Transformer is proposed to enhance local detail information. In the meanwhile, we propose a novel Transformer-based hierarchical discriminator to obtain diverse perspectives and lighting distribution knowledge. Comprehensive experiments on two challenging nighttime UAV benchmarks demonstrate that the proposed MVDANT achieves superior UAV tracking performance in both precision and efficiency. Quantitative tests in real-world settings fully prove the effectiveness of our work. The source code locates on https://github.com/vision4robotics/MVDANT.**

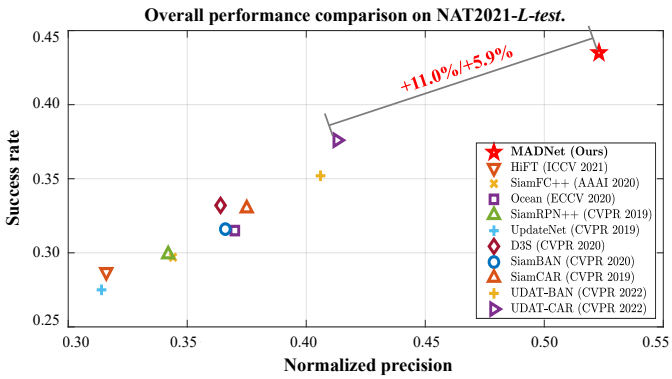

Fig. 1. The proposed MVDANT effectively adapted to nighttime aerial scenes from multiple views and yields favorable performance on NAT2021-*L-test*.

I. Introduction

Visual tracking is one of the most fundamental tasks in intelligent unmanned aerial systems, which aims to estimate the location of an object frame by frame given the initial state. This task is increasingly applied in autonomous landing [1], autonomous aerial manipulation operations [2], and self-localization [3]. Meanwhile, the significance of UAVs in low-light applications is increasing due to their distinct capabilities in hazardous or challenging environments. In the intervening years, various deep learning-based trackers [4]–[9] have continued to set state-of-the-arts (SOTAs) through large-scale benchmarks in bright light conditions. Although aerial tracking has made significant advances, the tracking performance of these trackers is severely suppressed in low-light conditions since a huge gap exists between daytime and nighttime scenarios, making the automation and applied range of UAV tracking still a formidable challenge.

Typically, nighttime scenes exhibit low illumination, high-level noise, and low contrast, making it difficult for general trackers trained with daytime images to effectively extract target features in low-light conditions. Unfortunately, few nighttime tracking benchmarks with sufficiently large and comprehensive annotations are available for direct training of nighttime trackers. Consequently, several studies [10]–[12]have developed low-light enhancers for pre-processing data in tracking pipelines. However, these methods lead to significant object information loss and limited adaptability to varying illumination conditions, resulting in over-enhancement and over-saturation for the bright regions and seriously degrading tracking performance in real-world scenarios. Accordingly, domain adaptation provides a novel solution for effectively improving the nighttime tracking performance of trackers by adapting to low-light conditions through implicit feature-level alignment [7], which extends the application of trackers in intelligent unmanned aerial systems.

Domain adaptation (DA) is the technique of fine-tuning a model that is initially trained on a source domain, to effectively generalize to a different target domain with distinct data distributions. Due to the increasing demand for intelligent unmanned systems to operate in real-world scenarios with varying illumination conditions, daytime-nighttime domain adaptation (DNDA) has gained significant attention in multiple fields. DNDA aims to adapt models trained on the daytime domain to perform well on the nighttime domain, thereby narrowing the gap between daytime and nighttime. However, existing methods only utilize knowledge from a single-view and cannot effectively handle annotated benchmarks from multiple observation views, resulting in biased predictions in the target domain with different observation angles. This poses significant challenges for aerial-view transformations during low-light UAV tracking.

The variation of viewpoints in UAV tracking, mainly attributed to target occlusion, limited flight trajectories, and rapid object movements, has a substantial impact on tracking performance due to the viewpoint and motion pattern disparities between multiple views. Moreover, shadow occlusion,

uneven lighting distribution, and disruptive noise exacerbate feature differences between multiple views in low-light conditions due to reduced object discriminability resulting from low illumination and the occurrence of severe noise or blur in the images. Although merging images from different capturing perspectives into a combined source domain is the most direct approach, it does not fully exploit abundant knowledge across multiple source domains, restricting their ability to learn more effective domain adaptation models. Some online tracking methods [13]–[16] utilize template updating to enhance robustness against viewpoint changes. However, these methods are vulnerable to accumulation errors and exhibit poor tracking performance in low-light conditions. Hence, it is crucial to develop robust tracking method that incorporate multi-view information under challenging low-light conditions.

In this work, a multi-view domain adaptation framework considering observation views for nighttime aerial tracking, namely MVDANT, is proposed to bridge the considerable gap between daytime and nighttime scenarios. As shown in Fig. 2, we capture images using UAVs at various flight altitudes and angles during the daytime. Additionally, a novel transformer feature alignment module considering multiple views is proposed to transform low-level features into high-level features with implicit multi-view information and semantic cues to improve feature extraction. Meanwhile, a Transformer-based hierarchical discriminator, with the ability to obtain diverse perspectives and lighting distribution is designed to facilitate aligning the source and target domain features. As shown in Fig. 1, the proposed MVDANT has achieved robust performance under multi-view aerial nighttime scenarios. The following are the main contributions of this work:

- A universal framework MVDANT considering perspective variation is proposed for nighttime aerial tracking to bridge the gap between the general daytime conditions and aerial nighttime conditions from multiple views.
- A novel multi-view transformer feature alignment module is proposed to align target domain at nighttime with source domains at daytime from multiple views to improve feature extraction.
- We introduce a Transformer-based hierarchical discriminator, which can capture diverse perspectives and lighting distribution knowledge to facilitate adversarial training in the nighttime.
- The nighttime tracking performance of MVDANT in comparison to other state-of-the-art (SOTA) trackers has been confirmed by a thorough analysis of public nighttime tracking benchmarks and a real-world test.

II. Related Work

A. Visual Tracking

Object tracking methods majorly comprise correlation filter-based methods and methods based on convolutional neural networks (CNNs). DCF-based trackers [17]–[19] were used in UAV tracking initially, because of their competitive and efficient performance while maintaining acceptable speed. However, complex optimization strategies have limited the development of DCF-based trackers despite their high performance. After SINT [20] modified the tracking task to patch matching, SiamFC [21] developed an end-to-end tracking method to discover similarities. SiamRPN [22] and SiamRPN++ [23] incorporate region suggestion networks (RPNs) into their Siamese-based framework. SiamFC++ [24] and SiamCAR [25], as solutions employing an anchor-free tracker, solve the aforementioned classification issue by adjusting the centroid and regressing on four offsets. Although Siamese networks have shown remarkable practicality and robustness for aerial tracking applications, few can guarantee superior performance under low illumination.

B. Nighttime aerial tracking

In low-light scenarios, the tracking performance is significantly reduced with low visibility and weak features. Several studies [11], [12] have developed tracking-related low-illumination enhancers for data pre-processing in the tracking pipeline. SCT [12] proposes a spatial-channel Transformer-based enhancer for low-light UAV tracking. HighlightNet [11] facilitates human perception and UAV tracking tasks through global feature modeling, pixel-level range masks, and a soft truncation mechanism. However, these enhancers lacks complexity and adaptability under varying illumination conditions, which may result in over-enhancement and over-saturation for the bright regions of the low-light images and UAV tracking failure. Therefore, other works [7] recommend employing a domain-adaptive strategy to bridge the gap between daytime and nighttime scenes, demonstrating robustness on public nighttime benchmarks.

C. Daytime-Nighttime Domain Adaptation

Domain adaptation has been applied to a variety of image classification scenario [26], [27] to reduce domain differences and transfer knowledge from the source domain to the target domain. In order to detect objects at night, Y.Sasagawa *et al.* [27] combined a low-light image enhancement model and an object detection model. In addition, adaptive techniques [28], [29] are also employed in semantic segmentation. Recently, an unsupervised domain adaptation framework [7] for object tracking has emerged. However, existing methods neglects the significant gap between multiple observation views. Transferring knowledge from different scenarios with only a single perspective is inadequate for addressing complex real-world scenarios.

III. Proposed Approach

In unsupervised MVDANT, the following scenario is examined: M labeled source domains, $S_1, S_2, ..., S_M$, acquired by UAV from various perspectives, and one unlabeled target domain T. In the i-th source domain, $S_i = \{(\mathbf{X}_{Si}^j, \mathbf{Z}_{Si}^j)\}_{j=1}^{N_i}$ suppose $\mathbf{X}_{Si}^j$ represents the search region and $\mathbf{Z}_{Si}^j$ represents the template patch in the j-th videos, where N_i is the number

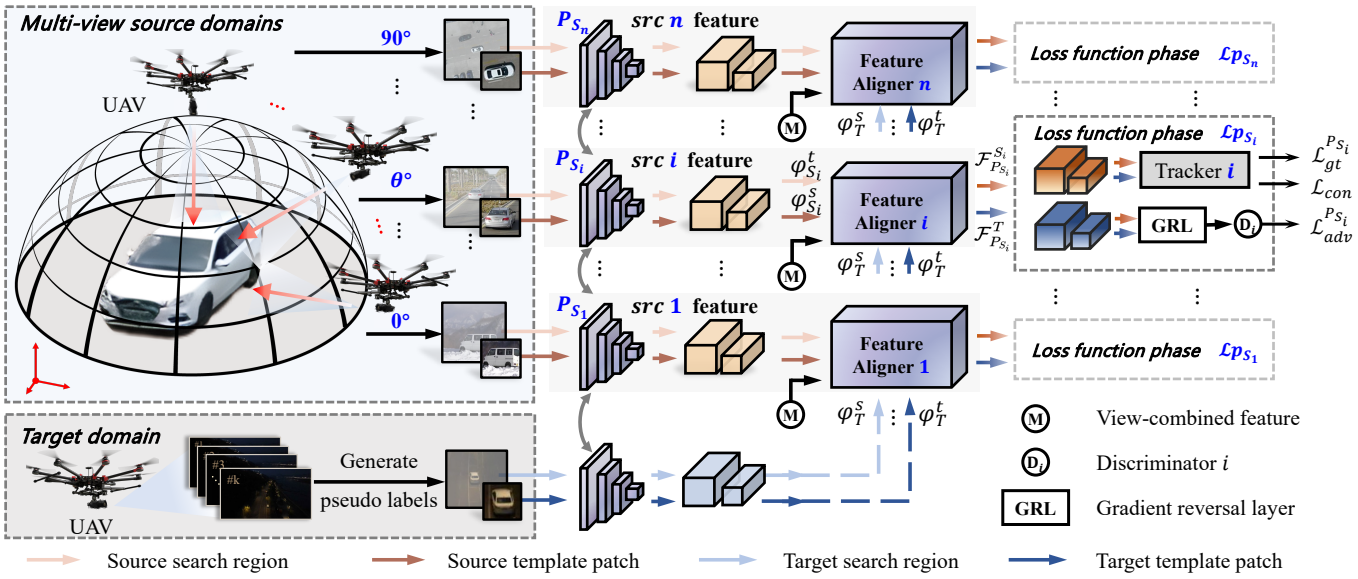

Fig. 2. Overview of MVDANT. The source domain data consists of the UAV captured in various daytime scene views as input, while the target domain is obtained from the UAV captured in the nighttime scene. *Feature extraction network*, *feature alignment network*, and *loss function phase* are the modules from left to right. The workflow of features from the target and source domains, respectively, is represented by two arrows of varying colors. Color saturation is used to differentiate the search domain from the template.

of videos in the i-th source domain. For the unlabeled target domain T, we use the common unsupervised data processing method to obtain potential targets and crop them. Consequently, the nighttime target domain can be represented by the set $T = \{(\mathbf{X}_t^j, \mathbf{Z}_t^j)\}_{j=1}^{N_T}$, suppose $\mathbf{X}_t^j$ represents the search region and $\mathbf{Z}_t^j$ represents the template patch in the j-th videos, where N_T is the number of target videos.

A novel end-to-end multi-source domain adaptive network called MVDANT is proposed for nighttime aerial tracking, and its pipeline is shown in Fig. 2.

A. Feature Alignment

Low-level features. Siamese network feature extraction comprises two branches: the template branch and the search branch. Search patches **X** and template patches **Z** corresponding to distinct source domains and the target domain are therefore simultaneously input into the network, and use a weight-sharing feature extractor to obtain the low-level feature map φ_T^s, φ_T^t in the target domain and $\varphi_{S_i}^s, \varphi_{S_i}^t$ in various source domains.

High-level features. The n tracking perspective branches, $P = \{P_{S_i}, i = 1, 2, \cdots, n\}$, are comprised of low-level feature maps of source domains from multiple views and the target domain. The low-level features from each viewpoint are input into the multi-view feature aligner $\mathcal{F}_{P_{S_i}}$ to improve feature extraction and generate high-level features.

Multi-view Feature Aligner. A multi-view transformer structure generates high-level features to facilitate feature extraction from multiple views. Multi-view feature encoder layer and current-view feature decoder layer are the principal components of the proposed feature alignment module.

Multi-view feature encoder layer seeks to identify the interdependencies between the target feature and information from multiple views. The multi-head attention module mAtt

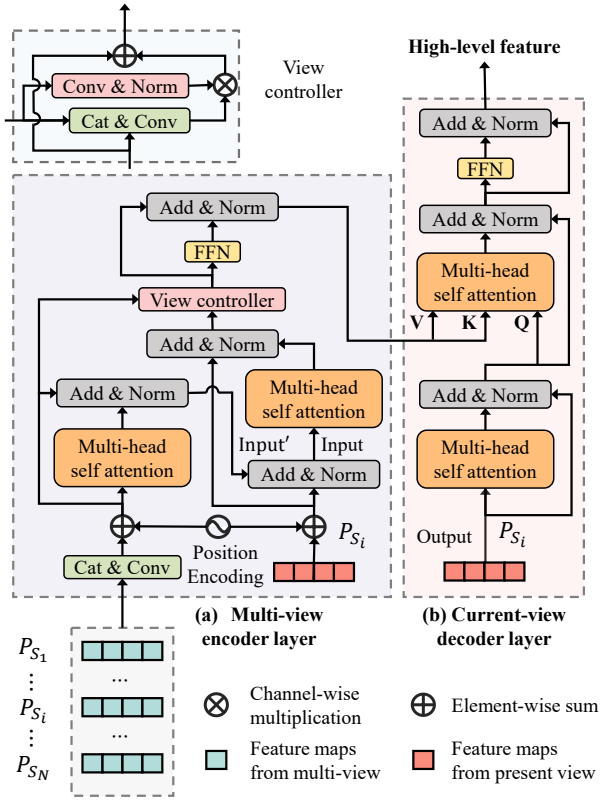

Fig. 3. Detailed workflow of the Multi-view feature aligner. The left sub-window displays the feature encoder and fuses the multi-source domain features; the right sub-window displays the decoder's structure and outputs the feature alignment results.

is formalized as follows [30]:

$$\begin{aligned} \mathrm{mAtt}(\mathbf{Q},\mathbf{K},\mathbf{V}) &= \left(\mathrm{Cat}\left(h^1,\ldots,h^N\right)\right)\mathbf{W}_c \quad , \\ h^j &= \mathrm{Att}\left(\mathbf{Q}\mathbf{W}_1^j,\mathbf{K}\mathbf{W}_2^j,\mathbf{V}\mathbf{W}_3^j\right) \quad , \end{aligned} \tag{1}$$

where N donates the number of parallel attention heads, and Att represents scaled dot-product attention.

To extract view-invariant features, features of the search patch and template patch are concatenated as view-combined features for the first multi-head attention module, and we take the instance of the template patch in the following introduction for clarity.

$$\begin{aligned} \mathbf{M} &= \mathrm{Conv}(\mathrm{Cat}(\mathcal{F}_{S_1}^t,\mathcal{F}_{S_2}^t,\ldots,\mathcal{F}_{S_n}^t) \quad , \\ \mathrm{Input}' &= \mathrm{Norm}(\mathrm{mAtt}(\mathbf{M}+\mathbf{P})+\mathbf{M}+\mathbf{P}) \quad , \end{aligned} \tag{2}$$

where $\mathbf{M}$ donates the view-combined feature, Conv represents the fractionally-strided convolutions, mAtt indicates the multi-head attention, $\mathbf{P}$ donates positional encodings, and Norm refers to the normalization layer.

The template features from present view $\mathcal{F}_{S_i}^t$, and input$'$ are input to the second multi-headed attention module.

$$\mathrm{Input} = \mathrm{Norm}(\mathcal{F}_{S_i}^t + \mathrm{Input}') \quad . \tag{3}$$

The output combines information about target features from the present view, with the structure of the decoder exhibited in Fig. 3. Thus, for each perspective branch P_{S_i}, such a loss function phase $\mathcal{L}_{P_{S_i}}$ can be constructed:

$$\mathcal{L}_{P_{S_i}} = \{(\mathcal{F}_{P_{S_i}}^{S_i},\mathcal{F}_{P_{S_i}}^{T}),\ i=1,2,\cdots,N\} \quad , \tag{4}$$

where $\mathcal{F}_{P_{S_i}}^{S_i}$ and $\mathcal{F}_{P_{S_i}}^{T}$ indicate the feature after alignment of the source domain data and the target domain data from the perspective P_{S_i}, respectively.

Remark 1: Through the multi-view feature alignment module, view-invariant features are enhanced with the view and semantic information in target features. Simultaneously, the view control layer aggregates inter-dependencies between various features, contributing to improving the robustness of tracking objects from diverse views.

B. Tracker Alignment

Discriminator in multi-view. For each perspective, daytime images are distinguished from nighttime images using discriminators, *i.e.,* $\mathrm{D} = \{\mathrm{D}_i, i=1,2,\cdots,n\}$, where D_i indicates the discriminator under perspective branch P_{S_i}. A gradient reversal layer (GRL) is placed between the feature aligner and domain discriminator to perform adversarial learning.

As the feature distribution varies from multiple perspectives, the discriminator for each perspective can be viewed as a subspace of the day-night feature space for discrimination. The high-level features are represented as:

$$\begin{aligned} \mathrm{D}_i\left(\mathcal{F}_{P_{S_j}}^{S_j}\right) &\rightarrow \mathrm{D}_i\left(S_j\right) \quad , \\ \mathrm{D}_i\left(\mathcal{F}_{P_{S_j}}^{T}\right) &\rightarrow \mathrm{D}_i\left(T_j\right) \quad . \end{aligned} \tag{5}$$

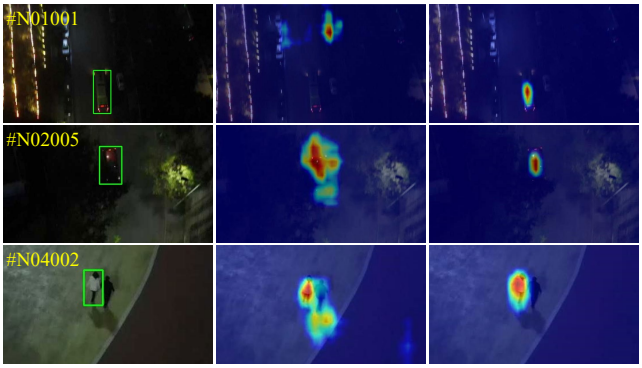

Fig. 4. Visual comparison of confidence maps generated by the baseline and the proposed MVDANT. Target objects are marked by green boxes.

Adversarial loss. In adversarial learning, a least-squares loss function is used to train the generator G to generate source domain features in target domain images and deceive the discriminator D at freezing to align the target domain with each source domain:

$$\mathcal{L}_{Adv} = \sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_{adv}^i\left(\mathrm{D}_i\left(T_j\right) - l_s\right) \quad , \tag{6}$$

where λ_{adv}^i represents the adversarial loss and ℓ_s denotes the label for the source domain.

Discriminator loss. Typically, the discriminator is implemented as a network, necessitating the learning of new parameters. The loss function of D is defined as:

(7)

where λ_d^i represents the discriminator loss, and ℓ_t denotes the label for the target domain.

Remark 2: In actual training, the daytime label of the source domain ℓ_s is assigned to 0 and the nighttime label of the target domain ℓ_t is assigned to 1.

Tracker consistency. An implicit strategy is employed to bridge the gap between each source and target in each tracking perspective. However, trackers trained on a single perspective tend to misidentify target images near the category boundary. Hence, the tracker's prediction results are regularized for the same target image in various loss function phases. The overall consistency cost is formalized as:

$$\mathcal{L}_{con} = \frac{2}{N(N-1)}\sum_{j=1}^{N-1}\sum_{i=j+1}^{N}\mathcal{T}_{\mathcal{M}}|B_{P_{S_i}}^T - B_{P_{S_j}}^T| \quad , \tag{8}$$

where $B_{P_{S_i}}^T$ denotes the bounding box prediction of the target image under perspective branch P_{S_i}, $\mathcal{T}_{\mathcal{M}}$ represents the mean squared error for various tracker calculation metrics.

Remark 3: Multiple tracker network architectures can be replaced, and SiamCAR [25] is adopted as the baseline tracker in this work, including classification, center-ness, and regression. In addition, the feature gap between multiple views in each daytime or nighttime scene is further narrowed.

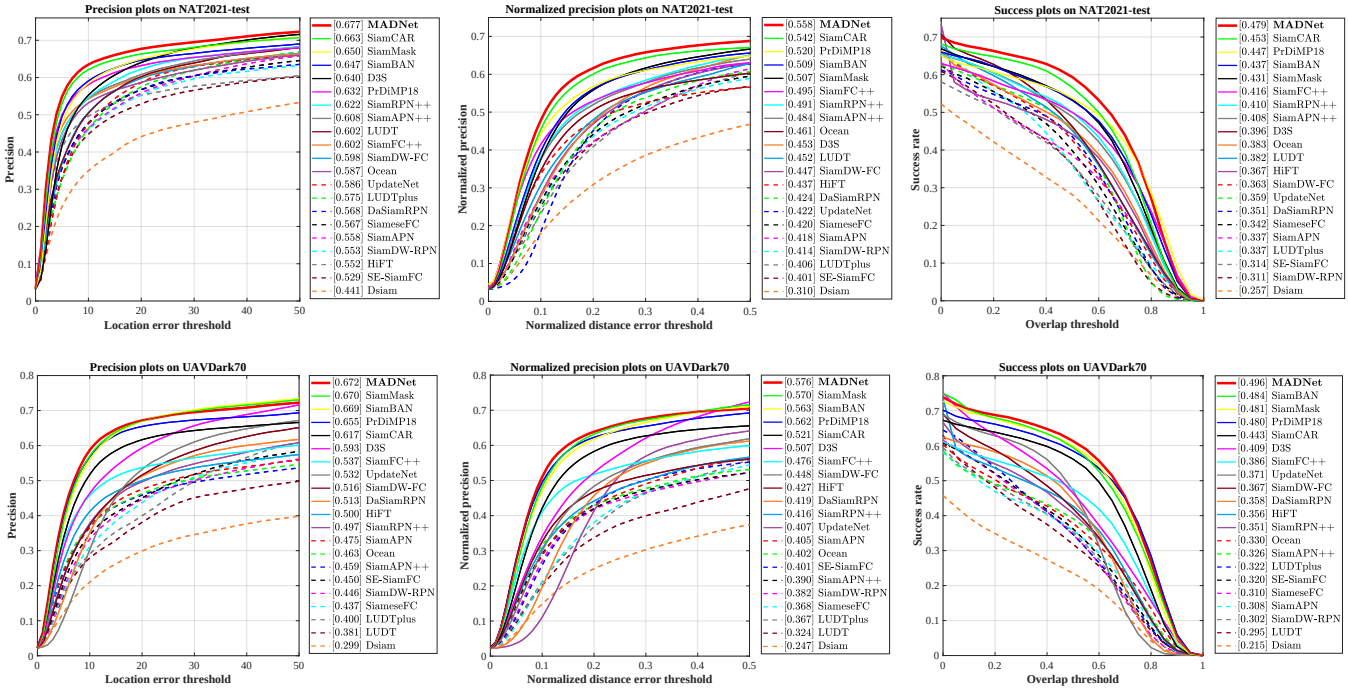

Fig. 5. Overall performance of SOTA trackers and MVDANT on nighttime aerial tracking benchmarks. The results show that the proposed MVDANT trackers realize top-ranked performance and improve baseline trackers favorably.

C. Overall Objective

The total training loss of the generator is summarized as follows:

$$\mathcal{L}_{Total} = \mathcal{L}_{GT} + \lambda\left(\mathcal{L}_{Adv} + \mathcal{L}_{con}\right) \quad , \tag{9}$$

where $\mathcal{L}_{GT}$ donates the classification and regression loss, and λ represents a weight to balance the loss from various views. During model training, set λ as 0.1 in implementation.

IV. Experiment

A. Experimental Setup

Dataset. Our framework is trained on five daytime public authoritative benchmarks.

- **GOT-10K** [31] is a large, high-diversity benchmark for general-purpose object tracking in the field including a total of 10,000 video clips of real-world moving objects.
- **NAT2021** [7] is a nighttime aerial tracking benchmark, and provides unlabeled nighttime tracking video for unsupervised training, which consists of 1400 videos containing over 276K frames,
- **UAV123** [32] is a drone-captured video tracking dataset containing over 110K frames and 123 video sequences, which has a pristine background.
- **UAVDT** [33] is a dataset comprised of approximately 8,000 frames with 14 attributes, based on vehicle traffic content ingested by UAVs. We use the UAV-benchmark-S In this work.
- **UAVTrack112** [34] is a benchmark which is created from images captured during real-world tests including 45 sequences.

In the daytime, UAV123, UAVDT, and UAVTrack112 are combined to simulate source domain data from a high-angle aerial view, while the GOT-10K is to simulate source domain data from a low-angle horizontal view. Additionally, NAT2021 is used as the target domain.

Remark 4: Since the UAVDT contains images collected under various weather conditions, we eliminated the training videos captured in low-light conditions to improve the distinction between the source and target domains.

Evaluation Metrics & Compared Baselines. To evaluate the impact of multi-source domain adaptation, the pre-trained tracking models are trained on different source domain benchmarks and employed as baseline models. In addition, We rank the performance in terms of success rate, precision, and normalized precision using a one-time evaluation (OPE).

B. Implementation Details.

Stage1. Source-model pre-training. In this stage, we separately pre-train the two source domains on the SiamCAR network, the batch size is set as 32 and a total of 20 epochs are performed by using stochastic gradient descent (SGD) with an initial learning rate of 0.001. The results in two pre-trained models are used as initialization weights for both trackers in MVDANT.

Stage2. Multi-source domain adaptation. We implement our MVDANT framework using PyTorch on an NVIDIA RTX A100 4 GPUs, and the discriminator is trained through the Adam optimizer. The base learning rate is set to 0.005 and decayed with a power of 0.8 according to the poly learning rate policy. The whole training process lasted for 20 epochs.

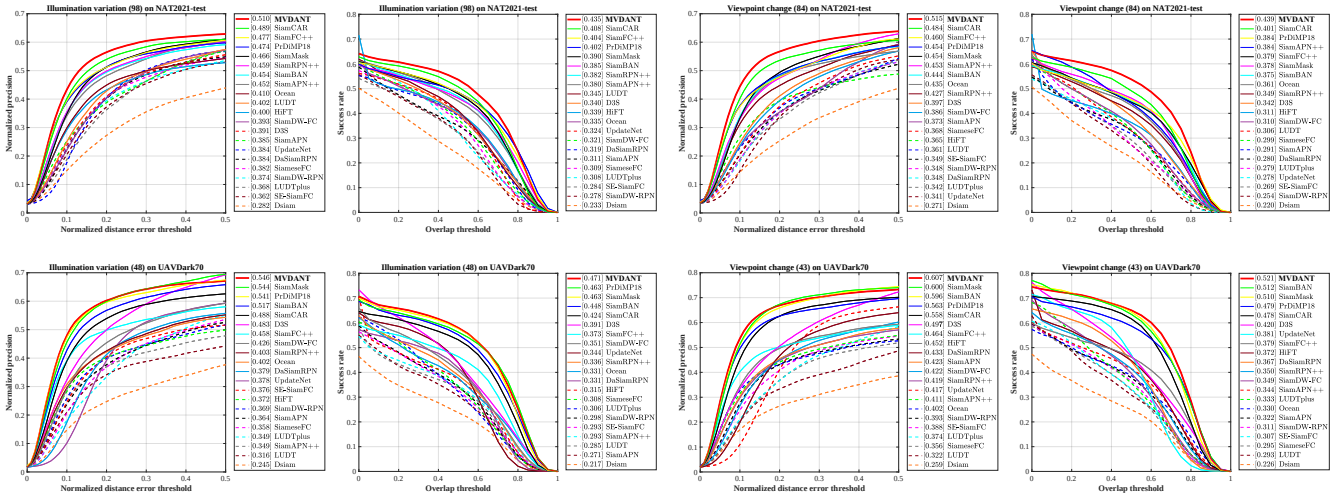

Fig. 6. Normalized precision plots and success plots of illumination and viewpoint attributes on NAT2021-*test* and UAVDark70.

TABLE I

THE RESULTS OF THE TOP EIGHT TRACKERS ON THE NAT2021-*L-test*. OUR TRACKER OUTPERFORMS ALL OTHER TRACKERS WITH AN OBVIOUS IMPROVEMENT. THE TOP THREE PERFORMANCES ARE RESPECTIVELY HIGHLIGHTED BY RED, GREEN, AND BLUE COLOR.

Trackers	SiamFC++ [24]	Ocean [35]	SiamRPN++ [23]	UpdateNet [36]	D3S [37]	UDAT-BAN [7]	UDAT-CAR [7]	**MVDANT**
Prec.	0.425	0.454	0.431	0.434	0.492	**0.496**	**0.506**	**0.577**
Norm. Prec.	0.344	0.370	0.342	0.314	0.364	**0.406**	**0.413**	**0.523**
Succ.	0.297	0.315	0.299	0.275	0.332	**0.352**	**0.376**	**0.435**

TABLE II

ADAPTIVE STRATEGIES FOR DIVERSE DATASETS. OUR TRACKER OUTPERFORMS ALL OTHER TRACKERS WITH AN OBVIOUS IMPROVEMENT. THE TOP THREE PERFORMANCES ARE RESPECTIVELY HIGHLIGHTED BY RED, GREEN, AND BLUE COLOR.

standards		NAT2021-*test*			UAVDark70		
		Prec.	Norm. Prec.	Succ.	Prec.	Norm. Prec.	Succ.
source-only [25]	Low-view-only	0.572	0.502	0.388	0.400	0.371	0.291
	High-view-only	0.518	0.415	0.333	0.242	0.226	0.174
	Source-combined	0.561	0.463	0.358	0.347	0.318	0.233
Single-source DA [7]	Low-view-only	**0.654**	**0.565**	**0.454**	**0.626**	**0.549**	**0.466**
	High-view-only	0.651	0.542	0.446	0.595	0.531	0.435
	Source-combined	**0.669**	**0.590**	**0.471**	**0.655**	**0.570**	**0.480**
Multi-source DA	\	**0.677**	**0.611**	**0.483**	**0.672**	**0.576**	**0.496**

C. Overall Performance

Comparison with SOTA Trackers. As shown in Fig. 5, MVDANT is **2.6**% better than SiamCAR (0.453) on NAT2021-*test* and **1.2**% better than SiamBAN (0.484) on UAVDark70 in success rate; in normalized precision, MVDANT is **1.6**% higher than SiamCAR (0.542) on NAT2021-*test* and **0.6**% higher than SiamMask (0.570) on UAVDark70. MVDANT trained on the preceding benchmarks achieves nighttime tracking performance comparable to that of other SOTA trackers.

Long-term tracking evaluation. To validate the effectiveness of our framework in long-term tracking performance, we conduct the evaluations on NAT2021-*L-test*. MVDANT outperformed the runner-up on the NAT2021-*L-test* by **7.1**% in precision, **11.0**% in normalized precision, and **5.9**% in success rate. The results presented in Table 2 demonstrate that MVDANT achieves highly competitive long-term tracking performance, significantly outperforming the baseline tracker.

D. Attribute-Based Performance

Additional environmental changes caused by illumination and views can exacerbate the difficulty of aerial tracking. To thoroughly assess the robustness of our tracker against particular challenges, a comparison of their pertinent properties is conducted, such as illumination variation, low resolution, fast motion, viewpoint change, *etc*. The comparison between other SOTA trackers is presented in Fig. 6. proves the robustness of our framework in several challenging conditions. For instance, MVDANT raises the success rate of the existing best performance by ∼**6.6**% on NAT2021-*test* for illumina-

tion variation. In addition, MVDANT realizes a success rate of 0.521 for viewpoint change on UAVDark70 and 0.476 for fast motion on the NAT2021-*test*, which improves the existing best performance by ∼**4.3**%.

E. Ablation Study

Effectiveness of MVDANT. In this work, MVDANT is compared to various benchmarks on the same training condition. Table 1 demonstrates that our model narrows the huge gap between the general daytime conditions and aerial nighttime conditions from multiple views. Specifically, MVDANT achieves significant improvement compared to the view-combined domain adaptation by ∼**3.5**% for Norm. Prec. and ∼**2.5**% for Succ. on NAT-*test*, while ∼**1.1**% for Norm. Prec. and ∼**3.3**% for Succ. on UAVDark70, respectively.

In addition, regardless of whether the adaptive method is employed or not, the performance is inferior when only the high-angle view is used as the source domain compared to the low-angle view, indicating that due to factors such as scale variation and low resolution, the high-angle view images captured by UAVs provides more background information and less about target local features.

Comparison with various modules activated. To investigate the performance of several MVDANT variations, ablation studies regarding various modules are presented in this subsection. This work considers `Baseline` as the model with SiamCAR with a ResNet50 backbone. `ADA` denotes adversarial multi-source domain adaptation. `MFA` represents a multi-view feature aligner with a novel transformer structure. `TA` denotes the method of tracker alignment. Table I contains the tracking results for NAT2021-*L-test*. The first row represents the original baseline, which demonstrates subpar performance. However, the addition of the entire MVDANT improved the Norm. Prec. and Succ. by **26.67**% and **32.01**%, respectively, demonstrating the effectiveness of the added modules.

F. Real-World Tests

MVDANT was implemented on a typical embedded system, the NVIDIA Jetson AGX Xavier, to demonstrate its applicability in nighttime drone tracking applications in the real world. Without TensorRT acceleration, MVDANT achieves an impressive real-time speed of 31.25 frames per second (FPS). In addition, Fig. 7 depicts nighttime tracking tests and CLE curves conducted in the real world. The CLE curves indicate that the prediction error is within 20 pixels, making tracking reliable. The real-world tests on our practical UAV strongly demonstrate the practicability and achieve robust nighttime object tracking of our proposed MVDANT.

TABLE III

MVDANT ON NAT2021-*L-test* COMPARISON WITH VARIOUS MODULES ACTIVATED. THE MOST ADVANTAGEOUS RESULTS ARE INDICATED IN RED. Δ INDICATES THE PERCENTAGE INCREASE OVER THE BASELINE.

ADA	MFA	TA	Norm. Prec.	Δ p(%)	Succ.	Δ s(%)
			0.375	-	0.330	–
✓			0.447	+19.20	0.362	+9.70
✓		✓	0.459	+22.40	0.381	+15.45
✓	✓		0.487	+29.87	0.410	+24.24
✓	✓	✓	**0.523**	**+39.47**	**0.435**	**+31.82**

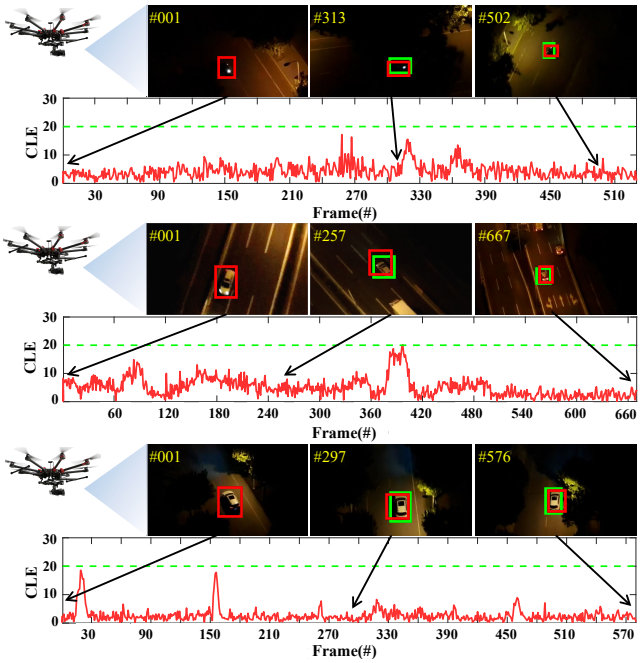

Fig. 7. Real-world tests on a typical UAV platform. Red bounding boxes denote the estimated positions. **CLE** curves between predictions and ground truth are drawn below. The green dashed line locates a threshold of 20 pixels, tracking errors within which are normally regarded as satisfying. The base tracker realizes favorable nighttime tracking assisted by MVDANT.

V. CONCLUSION

We propose using a multi-source domain adaptive approach MVDANT to address UAV nighttime tracking from multiple perspectives. A multi-source domain adaptive processing method is proposed to obtain high-level features by fusing the feature alignment network of multi-view features, aligning the daytime source domain from different capture views with the target domain of the night scene, and optimizing the loss function to align the tracker. With the same dataset training, the multi-source domain demonstrates a more effective structural advantage than other methods and has been tested on several publicly available datasets, particularly for Long-term tracking, where MVDANT demonstrates a very stable tracking performance.

VI. ACKNOWLEDGEMENT

This work is supported by the Natural Science Foundation of Shanghai (No. 20ZR1460100) and the National Natural Science Foundation of China (No. 62173249).

REFERENCES

[1] G. Niu, Q. Yang, Y. Gao, and M.-O. Pun, "Vision-Based Autonomous Landing for Unmanned Aerial and Ground Vehicles Cooperative Systems," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6234–6241, 2022.

[2] D. R. McArthur, Z. An, and D. J. Cappelleri, "Pose-Estimate-Based Target Tracking for Human-Guided Remote Sensor Mounting with a UAV," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 10 636–10 642.

[3] J. Ye, C. Fu, F. Lin, F. Ding, S. An, and G. Lu, "Multi-Regularized Correlation Filter for UAV Tracking and Self-Localization," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 6, pp. 6004–6014, 2022.

[4] C. Fu, K. Lu, G. Zheng, J. Ye, Z. Cao, and B. Li, "Siamese Object Tracking for Unmanned Aerial Vehicle: A Review and Comprehensive Analysis," *arXiv preprint arXiv:2205.04281*, 2022.

[5] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, "Transformer Tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8126–8135.

[6] Z. Cao, C. Fu, J. Ye, B. Li, and Y. Li, "HiFT: Hierarchical Feature Transformer for Aerial Tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15 457–15 466.

[7] J. Ye, C. Fu, G. Zheng, D. P. Paudel, and G. Chen, "Unsupervised Domain Adaptation for Nighttime Aerial Tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 8896–8905.

[8] Z. Cao, Z. Huang, L. Pan, S. Zhang, Z. Liu, and C. Fu, "TCTrack: Temporal Contexts for Aerial Tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14 798–14 808.

[9] Z. Cao, C. Fu, J. Ye, B. Li, and Y. Li, "SiamAPN++: Siamese Attentional Aggregation Network for Real-time Uav Tracking," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 3086–3092.

[10] J. Ye, C. Fu, G. Zheng, Z. Cao, and B. Li, "DarkLighter: Light Up the Darkness for UAV Tracking," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 3079–3085.

[11] C. Fu, H. Dong, J. Ye, G. Zheng, S. Li, and J. Zhao, "Highlightnet: Highlighting low-light potential features for real-time uav tracking," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 12 146–12 153.

[12] J. Ye, C. Fu, Z. Cao, S. An, G. Zheng, and B. Li, "Tracker Meets Night: A Transformer Enhancer for UAV Tracking," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3866–3873, 2022.

[13] N. Wang, W. Zhou, J. Wang, and H. Li, "Transformer meets tracker: Exploiting temporal context for robust visual tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1571–1580.

[14] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang, "Learning dynamic siamese network for visual object tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1763–1771.

[15] Z. Fu, Q. Liu, Z. Fu, and Y. Wang, "Stmtrack: Template-free visual tracking with space-time memory networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 774–13 783.

[16] B. Yan, H. Peng, J. Fu, D. Wang, and H. Lu, "Learning spatio-temporal transformer for visual tracking," in *IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 10 448–10 457.

[17] J. Ye, C. Fu, F. Lin, F. Ding, S. An, and G. Lu, "Multi-Regularized Correlation Filter for UAV Tracking and Self-Localization," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 6, pp. 6004–6014, 2022.

[18] G. Zheng, C. Fu, J. Ye, F. Lin, and F. Ding, "Mutation Sensitive Correlation Filter for Real-Time UAV Tracking with Adaptive Hybrid Label," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 503–509.

[19] Y. Li, C. Fu, F. Ding, Z. Huang, and G. Lu, "AutoTrack: Towards High-performance Visual Tracking for UAV with Automatic Spatio-temporal Regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 923–11 932.

[20] R. Tao, E. Gavves, and A. W. Smeulders, "Siamese Instance Search for Tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1420–1429.

[21] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional Siamese Networks for Object Tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 850–865.

[22] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High Performance Visual Tracking with Siamese Region Proposal Network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8971–8980.

[23] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN++: Evolution of Siamese Visual Tracking with Very Deep Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4282–4291.

[24] Y. Xu, Z. Wang, Z. Li, Y. Yuan, and G. Yu, "SiamFC++: Towards Robust and Accurate Visual Tracking with Target Estimation Guidelines," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 34, no. 07, 2020, pp. 12 549–12 556.

[25] D. Guo, J. Wang, Y. Cui, Z. Wang, and S. Chen, "SiamCAR: Siamese Fully Convolutional Classification and Regression for Visual Tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6269–6277.

[26] P. Panareda Busto and J. Gall, "Open Set Domain Adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 754–763.

[27] W. Li, Z. Xu, D. Xu, D. Dai, and L. Van Gool, "Domain Generalization and Adaptation Using Low Rank Exemplar SVMs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1114–1127, 2017.

[28] X. Wu, Z. Wu, H. Guo, L. Ju, and S. Wang, "DANNet: A One-stage Domain Adaptation Network for Unsupervised Nighttime Semantic Segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15 769–15 778.

[29] M. Kim, S. Joung, S. Kim, J. Park, I.-J. Kim, and K. Sohn, "Cross-domain Grouping and Alignment for Domain Adaptive Semantic Segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021, pp. 1799–1807.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," *Advances in neural information processing systems*, vol. 30, 2017.

[31] L. Huang, X. Zhao, and K. Huang, "Got-10k: A Large High-diversity Benchmark for Generic Object Tracking in the Wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 5, pp. 1562–1577, 2019.

[32] M. Mueller, N. Smith, and B. Ghanem, "A Benchmark and Simulator for UAV Tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 445–461.

[33] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, "The Unmanned Aerial Vehicle Benchmark: Object Detection and Tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 370–386.

[34] C. Fu, Z. Cao, Y. Li, J. Ye, and C. Feng, "Siamese Anchor Proposal Network for High-Speed Aerial Tracking," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 1–7.

[35] Z. Zhang, H. Peng, J. Fu, B. Li, and W. Hu, "Ocean: Object-aware Anchor-free Tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 771–787.

[36] L. Zhang, A. Gonzalez-Garcia, J. v. d. Weijer, M. Danelljan, and F. S. Khan, "Learning the model update for siamese trackers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4010–4019.

[37] A. Lukezic, J. Matas, and M. Kristan, "D3S-a Discriminative Single Shot Segmentation Tracker," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 7133–7142.